On learning to localize objects with minimal supervision

Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, Trevor Darrell

On learning to localize objects with minimal supervision

Hyun Oh Song SONG@EECS.BERKELEY.EDU
Ross Girshick RBG@EECS.BERKELEY.EDU
Stefanie Jegelka STEFJE@EECS.BERKELEY.EDU
Julien Mairal JULIEN.MAIRAL@INRIA.FR
Zaid Harchaoui ZAID.HARCHAOUI@INRIA.FR
Trevor Darrell TREVOR@EECS.BERKELEY.EDU

Abstract

Learning to localize objects with minimal supervision is an important problem in computer vision, since large fully annotated datasets are extremely costly to obtain. In this paper, we propose a new method that achieves this goal with only image-level labels of whether the objects are present or not. Our approach combines a discriminative submodular cover problem for automatically discovering a set of positive object windows with a smoothed latent SVM formulation. The latter allows us to leverage efficient quasi-Newton optimization techniques. Our experiments demonstrate that the proposed approach provides a 50% relative improvement in mean average precision over the current state-of-the-art on PASCAL VOC 2007 detection.

1. Introduction

The classical paradigm for learning object detection models starts by annotating each object instance, in all training images, with a bounding box. However, this exhaustive labeling approach is costly and error prone for large-scale datasets. The massive amount of textually annotated visual data available online inspires a different, more challenging, research problem. Can weakly-labeled imagery, without bounding boxes, be used to reliably train object detectors?

In this alternative paradigm, the goal is to learn to localize objects with minimal supervision (Weber et al., 2000a;b). We focus on the case where the learner has access to binary image labels that encode whether an image contains the target object or not, without access to any instance level annotations (i.e., bounding boxes).

Our approach starts by reducing the set of possible image locations that contain the object of interest from millions to thousands per image, using the selective search window proposal technique introduced by Uijlings et al. (2013). Then, we formulate a discriminative submodular cover algorithm to discover an initial set of image windows that are likely to contain the target object. After training a detection model with this initial set, we refine the detector using a novel smoothed formulation of latent SVM (Andrews et al., 2003; Felzenszwalb et al., 2010). We employ recently introduced object detection features, based on deep convolutional neural networks (Donahue et al., 2014; Girshick et al., 2014), to represent the window proposals for clustering and detector training.

Compared to prior work on weakly-supervised detector training, we show substantial improvements on the standard evaluation metric (detection average precision on PASCAL VOC). Quantitatively, our approach achieves a 50% relative improvement in mean average precision over the current state-of-the-art for weakly-supervised learning.

2. Related work

Our work is related to three active research areas: (1) weakly-supervised learning, (2) unsupervised discovery of mid-level visual elements, and (3) co-segmentation.

We build on a number of previous approaches for training object detectors from weakly-labeled data. In nearly all cases, the task is formulated as a multiple instance learning (MIL) problem (Long & Tan, 1996). In this formulation, the learner has access to an image-level label indicating the presence or absence of the target class, but not its location (if it is present). The challenge faced by the learner is to find the sliver of signal present in the positive images, but absent from the negative images. The implicit assumption is that this signal will correspond to the positive class.

Although there have been recent works on convex relaxations (Li et al., 2013; Joulin & Bach, 2012), most MIL algorithms start from an initialization and then perform some form of local optimization. Early efforts, such as (Weber

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

et al., 2000a;b; Galleguillos et al., 2008; Fergus et al., 2007; Crandall & Huttenlocher, 2006; Chum & Zisserman, 2007; Chen et al., 2013), focused on datasets with strong object-in-the-center biases (e.g. Caltech-101). This simplified setting enabled clarity and focus on the MIL formulation, image features, and classifier design, but masked the vexing problem of finding a good initialization in data where such helpful biases are absent.

More recent work, such as (Siva & Xiang, 2011; Siva et al., 2012), attempts to learn detectors, or simply automatically generate bounding box annotations from much more challenging datasets such as PASCAL VOC (Everingham et al., 2010). In this data regime, focusing on initialization is crucial and carefully designed heuristics, such as shrinking bounding boxes (Russakovsky et al., 2012), are often employed.

Recent literature on unsupervised mid-level visual element discovery (Doersch et al., 2012; Singh et al., 2012; Endres et al., 2013; Juneja et al., 2013; Raptis et al., 2012) uses weak labels to discover visual elements that occur commonly in positive images but not in negative images. Discovered visual element representation were shown to successfully provide discriminative information in classifying images into scene types. The most recent work (Doersch et al., 2013) presents a discriminative mode seeking formulation and draws connections between discovery and mean-shift algorithms (Fukunaga & Hostetler, 1975).

The problem of finding common structure is related to the challenging setting of co-segmentation (Rother et al., 2006; Joulin et al., 2010; Alexe et al., 2010), which is the unsupervised segmentation of an object that is present in multiple images. While in this paper we do not address pixel-level segmentation, we employ ideas from co-segmentation: the intuition behind our submodular cover framework in Section 4 is shared with CoSand (Kim et al., 2011). Finally, submodular covering ideas have recently been applied to (active) filtering of hypothesis after running a detector, and without the discriminative flavor we propose (Barinova et al., 2012; Chen et al., 2014).

3. Problem formulation

Our goal is to learn a detector for a visual category from a set of images, each with a binary label. We model an image as a set of overlapping rectangular windows and follow a standard approach to detection: reduce the problem of detection to the problem of binary classification of image windows. However, at training time we are only given image-level labels, which leads to a classic multiple instance learning (MIL) problem. We can think of each image as a "bag" of instances (rectangular windows) and the binary image label $y = 1$ specifies that the bag contains at least one instance of the target category. The label $y = -1$ specifies that the image contains no instances of the category. During training, no instance labels are available.

MIL problems are typically solved (locally) by finding a local minimum of a non-convex objective function, such as MI-SVM (Andrews et al., 2003). In practice, the quality of the local solution depends heavily on the quality of the initialization. We therefore focus extensively on finding a good initialization. In Section 4, we develop an initialization method by formulating a discriminative set multicover problem that can be solved approximately with a greedy algorithm. This initialization, without further MIL refinement, already produces good object detectors, validating our approach. However, we can further improve these detectors by optimizing the MIL objective. We explore two alternative MIL objectives in Section 5. The first is the standard Latent SVM (equivalently MI-SVM) objective function, which can be optimized by coordinate descent on an auxiliary objective that upper-bounds the LSVM objective. The second method is a novel technique that smoothes the Latent SVM objective and can be solved more directly with unconstrained smooth optimization techniques, such as L-BFGS (Nocedal & Wright, 1999). Our experimental results show modest improvements from our smoothed LSVM formulation on a variety of MIL datasets.

4. Finding objects via submodular cover

Learning with LSVM is a chicken and egg problem: The model weights are needed to infer latent annotations, but the latent annotations are needed to estimate the model weights. To initialize this process, we approximately identifying jointly present objects in a weakly supervised manner. The experiments show a significant effect from this initialization. Our procedure implements two essential assumptions: (i) the correct boxes are similar, in an appropriate feature space, across positive images (or there are few modes), and (ii) the correct boxes do not occur in the negative images. In short, in the similarity graph of all boxes we seek dense subgraphs that only span the positive images. Finding such subgraphs is a nontrivial combinatorial optimization problem.

The problem of finding and encoding a jointly present signal in images is an old one, and has been addressed by clustering, minimum description length priors, and the concept of exemplar (Darrell et al., 1990; Leibe et al., 2004; Mikolajczyk et al., 2006; Kim et al., 2011). These approaches share the idea that a small number of exemplars or clusters should well encode the shared information we are interested in. We formalize this intuition as a flexible *submodular cover* problem. However, we also have label information at hand that can help identify correct boxes. We therefore integrate into our covering framework the relevance

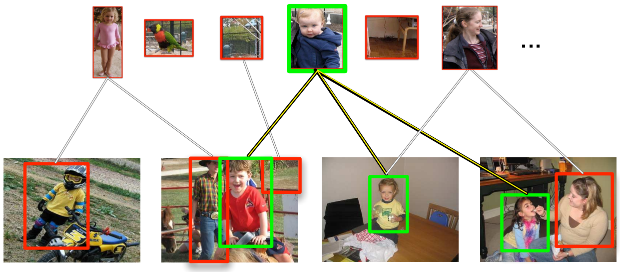

Figure 1. Illustration of the graph $\mathcal{G}$ with $\mathcal{V}$ (top row) and $\mathcal{U}$ (bottom row). Each box $b \in \mathcal{V}$ is connected to its closest neighbors from positive images (one from each image). Non-discriminative boxes occur in all images equally, and may not even have any boxes from positive images among their closest neighbors – and consequently no connections to $\mathcal{U}$. Picking the green-framed box v in $\mathcal{V}$ "covers" its (green) highlighted neighbors $\Gamma(b)$.

for positively versus negatively labeled images, generalizing ideas from (Doersch et al., 2012). This combination allows us to find multiple modes of the object appearance distribution.

Let $\mathcal{P}$ be the set of all positive images. Each image contains a set $\mathcal{B}_I = \{b_1, \ldots, b_m\}$ of candidate bounding boxes generated from selective search region proposals (Uijlings et al., 2013). In practice, there are about 2000 region proposal boxes per image and about 5000 training images in the PASCAL VOC dataset. Ultimately, we will define a function $F(S)$ on sets S of boxes that measures how well the set S represents $\mathcal{P}$. For each box b, we find its nearest neighbor box in each (positive *and negative*) image. We sort the set $\mathcal{N}(b)$ of all such neighbors of b in increasing order by their distance to b. This can be done in parallel. We will define a graph using these nearest neighbors that allows us to optimize for a small set of boxes S that are (i) *relevant* (occur in many positive images); (ii) *discriminative* (dissimilar to the boxes in the negative images); and (iii) *complementary* (capture multiple modes).

We construct a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{U}, \mathcal{E})$ whose nodes $\mathcal{V}$ and $\mathcal{U}$ are all boxes occurring in $\mathcal{P}$ (each b occurs once in $\mathcal{V}$ and once in $\mathcal{U}$). The nodes in $\mathcal{U}$ are partitioned into groups $\mathcal{B}_I$: $\mathcal{B}_I$ contains all boxes from image $I \in \mathcal{P}$. The edges $\mathcal{E}$ are formed by connecting each node (box) $b \in \mathcal{V}$ to its top k neighbors in $\mathcal{N}(b) \subseteq \mathcal{U}$ from positive images. Figure 1 illustrates the graph. Connecting only to the top k neighbors (instead of all) implements discriminativeness: the neighbors must compete. If b occurs in positively and negatively labeled images equally, then many top-k closest neighbors in $\mathcal{N}(b)$ stem from negative images. Consequently, b will not be connected to many nodes (boxes from $\mathcal{P}$) in $\mathcal{G}$. We denote the neighborhood of a set of nodes $S \subseteq \mathcal{V}$ by $\Gamma(S) = \{b \in \mathcal{U} \mid \exists (v, b) \in \mathcal{E} \text{ with } v \in S\}$.

Let $S \subseteq \mathcal{V}$ denote a set of selected boxes. We define a *covering score* $\mathrm{cov}_{I,t}(S)$ for each I that is determined by a *covering threshold* t and a scalar, nondecreasing concave function $g : \mathbb{R}_+ \to \mathbb{R}_+$:

$$\mathrm{cov}_{I,t}(S) = g(\min\{t,\ |\Gamma(S) \cap \mathcal{B}_I|\}). \tag{1}$$

This score measures how many boxes in $\mathcal{B}_I$ are neighbors of S and thus "covered". We gain from covering up to t boxes from $\mathcal{B}_I$ – anything beyond that is considered redundant. The *total covering score* of a set $S \subseteq \mathcal{V}$ is then

$$F(S) = \sum\nolimits_{I \in \mathcal{P}} \mathrm{cov}_{I,t}(S). \tag{2}$$

The threshold t balances relevance and complementarity: let, for simplicity, $g = \mathrm{id}$. If $t = 1$, then a set that maximizes $\mathrm{cov}_{I,t}(S)$ contains boxes from many different images, and few from a single image. The selected neighborhoods are very complementary, but some of them may not be very relevant and cover outliers. If t is large, then any additionally covered box yields a gain, and the best boxes $b \in \mathcal{V}$ are those with the largest degree. A box has large degree if many of its closest neighbors in $\mathcal{N}(b)$ are from positive images. This also means b is discriminative and relevant for $\mathcal{P}$.

Lemma 1. *The function $F : 2^{\mathcal{V}} \to \mathbb{R}_+$ defined in Equation* (2) *is nondecreasing and submodular.*

A set function is *submodular* if it satisfies *diminishing marginal returns*: for all v and $S \subseteq T \subseteq \mathcal{V} \setminus \{v\}$, it holds that $F(S \cup \{v\}) - F(S) \geq F(T \cup \{v\}) - F(T)$.

Proof. First, the function $S \mapsto |\Gamma(S) \cap \mathcal{B}_I|$ is a covering function and thus submodular: let $S \subset T \subseteq \mathcal{V} \setminus b$. Then $\Gamma(S) \subseteq \Gamma(T)$ and therefore

$$\begin{aligned} |\Gamma(T \cup \{b\})| - |\Gamma(T)| &= |\Gamma(b) \setminus \Gamma(T)| && (3)\\ &\leq |\Gamma(b) \setminus \Gamma(S)| && (4)\\ &= |\Gamma(S \cup \{b\})| - |\Gamma(S)|. && (5) \end{aligned}$$

The same holds when intersecting with $\mathcal{B}_I$. Thus, $\mathrm{cov}_{t,I}(S)$ is a nondecreasing concave function of a submodular function and therefore submodular. Finally, F is a sum of submodular functions and hence also submodular. Monotonicity is obvious. □

We aim to select a representative subset $S \subseteq \mathcal{V}$ with minimum cardinality:

$$\min_{S \subseteq \mathcal{V}} |S| \quad \text{s.t. } F(S) \geq \alpha F(\mathcal{V}) \tag{6}$$

for $\alpha \in (0, 1]$. We optimize this via a greedy algorithm: let $S_0 = \emptyset$ and, in each step τ, add the node v that maximizes the marginal gain $F(S_\tau \cup \{v\}) - F(S_\tau)$.

Lemma 2. *The greedy algorithm solves Problem* (6) *within an approximation factor of* $1 + \log\left(\frac{kg(1)}{g(t) - g(t-1)}\right) = O(\log k)$.

Lemma 2 says that the algorithm returns a set $\widehat{S}$ with $F(\widehat{S}) \geq \alpha F(\mathcal{V})$ and $|\widehat{S}| \leq O(\log k)|S^*|$, where S^* is an optimal solution. This result follows from the analysis by Wolsey (1982) (Thm. 1) adapted to our setting. To get a better intuition for the formulation (6) we list some special cases:

Min-cost cover. With $t = 1$ and $g(a) = a$ being the identity, Problem 6 becomes a min-cost cover problem. Such straightforward covering formulations have been used for filtering after running a detector (Barinova et al., 2012).

Maximum relevance. A minimum-cost cover merely focuses on complementarity of the selected nodes S, which may include rare outliers. At the other extreme (t large), we would merely select by the number of neighbors (Doersch et al. (2012) choose one single $\mathcal{N}(b)$ that way).

Multi-cover. To smoothly move between the two extremes, one may choose $t > 1$ and g to be sub-linear. This trades off representation, relevance, and discriminativeness.

In Figure 2, we visualize top 5 nearest neighbors with positive labels in the first chosen cluster S_1 for all 20 classes on the PASCAL VOC data. Our experiments in Section 6 show the benefits of our framework. Potentially, the results might improve even further when using the complementary mode shifts of (Doersch et al., 2013) as a pre-selection step before covering.

5. Iterative refinement with latent variables

In this section, we review the latent SVM formulation, and we propose a simple smoothing technique enabling us to use classical techniques for unconstrained smooth optimization. Figure 3 illustrates our multiple instance learning analogy for object detection with one-bit labels.

5.1. Review of latent SVM

For a binary classification problem, the latent SVM formulation consists of learning a decision function involving a maximization step over a discrete set of configurations $\mathcal{Z}$. Given a data point $\mathbf{x}$ in $\mathbb{R}^p$ that we want to classify, and some learned model parameters $\mathbf{w}$ in $\mathbb{R}^d$, we select a label y in $\{-1, +1\}$ as follows:

$$y = \operatorname{sign}\left(\max_{\mathbf{z}\in\mathcal{Z}} \mathbf{w}^\intercal \phi(\mathbf{x}, \mathbf{z})\right), \tag{7}$$

where $\mathbf{z}$ is called a "latent variable" chosen among the set $\mathcal{Z}$. For object detection, $\mathcal{Z}$ is typically a set of bounding boxes, and maximizing over $\mathcal{Z}$ amounts to finding a bounding box containing the object. In deformable part models (Felzenszwalb et al., 2010), the set $\mathcal{Z}$ contains all possible part configurations, each part being associated to a position in the image. The resulting set $\mathcal{Z}$ has exponential size, but (7) can be solved efficiently with dynamic programming techniques for particular choices of ϕ.

Learning the model parameters $\mathbf{w}$ is more involved than solving a simple SVM problem. We are given some training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where the vectors $\mathbf{x}_i$ are in $\mathbb{R}^p$ and the scalars y_i are binary labels in $\{1, -1\}$. Then, the latent SVM formulation becomes

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^n \ell\left(y_i,\ \max_{\mathbf{z}\in\mathcal{Z}} \mathbf{w}^\intercal\phi(\mathbf{x}_i, \mathbf{z})\right), \tag{8}$$

where $\ell : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is the hinge loss defined as $\ell(y, \hat{y}) = \max(0, 1-y\hat{y})$, which encourages the decision function for each training example to be the same as the corresponding label. Similarly, other loss functions can be used such as the logistic or squared hinge loss.

Problem (8) is nonconvex and nonsmooth, making it hard to tackle. A classical technique to obtain an approximate solution is to use a difference of convex (DC) programming technique, called concave-convex procedure (Yuille & Rangarajan, 2003; Yu & Joachims, 2009). We remark that the part of (8) corresponding to negative examples is convex with respect to $\mathbf{w}$. It is indeed easy to show that each corresponding term can be written as a pointwise maximum of convex functions, and is thus convex (see Boyd & Vandenberghe, 2004): when $y_i = -1$, $\ell\left(y_i, \max_{\mathbf{z}\in\mathcal{Z}} \mathbf{w}^\intercal\phi(\mathbf{x}_i, \mathbf{z})\right) = \max_{\mathbf{z}\in\mathcal{Z}} \ell(y_i, \mathbf{w}^\intercal\phi(\mathbf{x}_i, \mathbf{x}))$. On the other hand, the part corresponding to positive examples is concave, making the objective (8) suitable to DC programming. Even though such a procedure does not have any theoretical guarantee about the quality of the optimization, it monotonically decreases the value of the objective and performs relatively well when the problem is well initialized (Felzenszwalb et al., 2010).

We propose a *smooth formulation* of latent SVM, with two main motives. First, smoothing the objective function of latent SVM allows the use of efficient second-order optimization algorithms such as quasi-Newton (Nocedal & Wright, 1999) that can leverage curvature information to speed up convergence. Second, as we show later, smoothing the latent SVM boils down to considering the top-N configurations in the maximization step in place of the top-1 configuration in the regular latent SVM. As a result, the smooth latent SVM training becomes more robust to unreliable configurations in the early stages, since a larger set of plausible configurations is considered at each maximization step.

5.2. Smooth formulation of LSVM

In the objective (8), the hinge loss can be easily replaced by a smooth alternative, e.g., squared hinge, or logistic loss. However, the non-smooth points induced by the following functions are more difficult to handle

$$f_{\mathbf{x}_i}(\mathbf{w}) := \max_{\mathbf{z}\in\mathcal{Z}} \mathbf{w}^\intercal\phi(\mathbf{x}_i, \mathbf{z}). \tag{9}$$

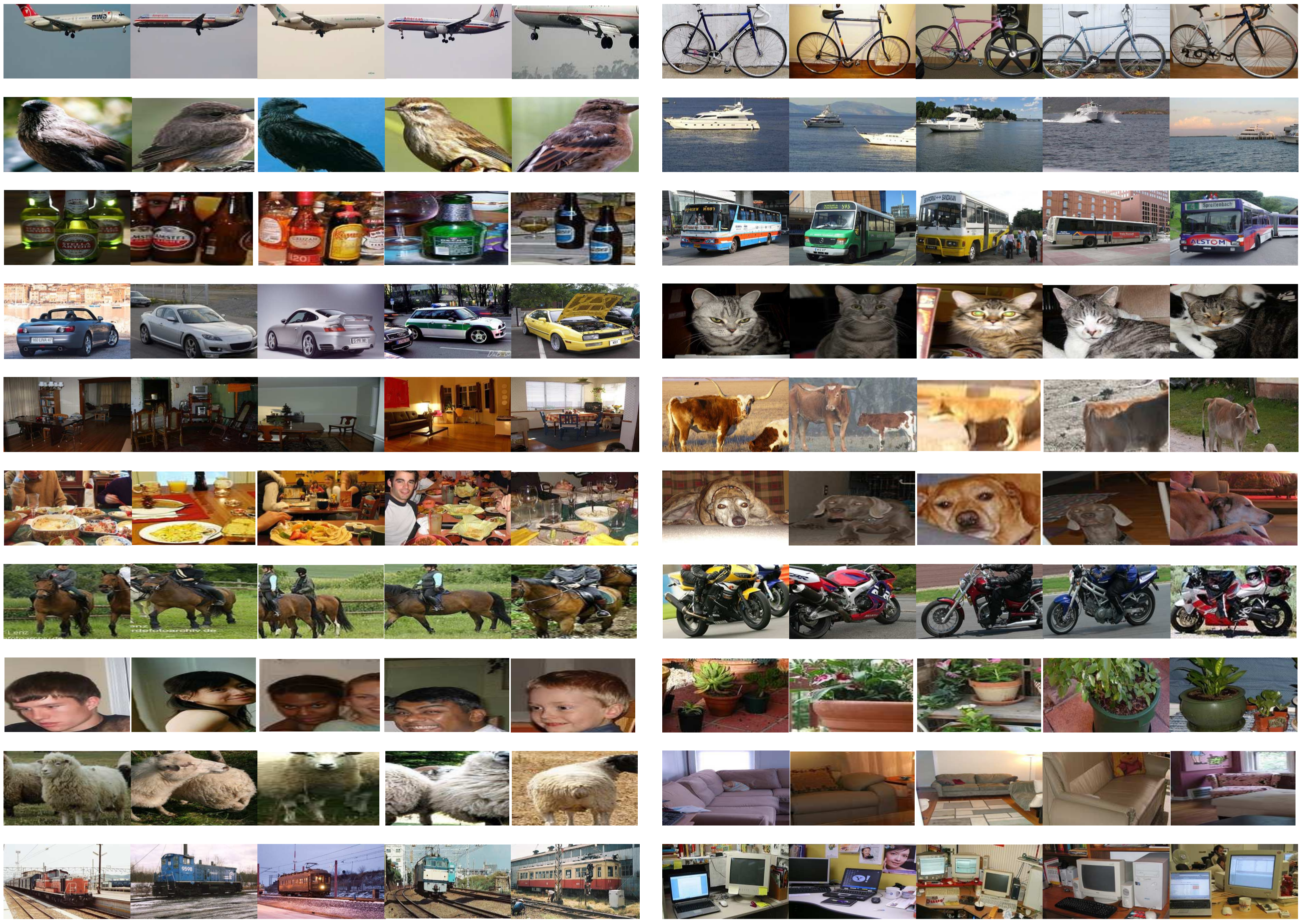

Figure 2. Visualizations of top 5 nearest neighbor proposal boxes with positive labels in the first cluster, S_1 for all 20 classes in PASCAL VOC dataset. From left to right, aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, diningtable, dog, horse, motorbike, person, plant, sheep, sofa, train, and tvmonitor.

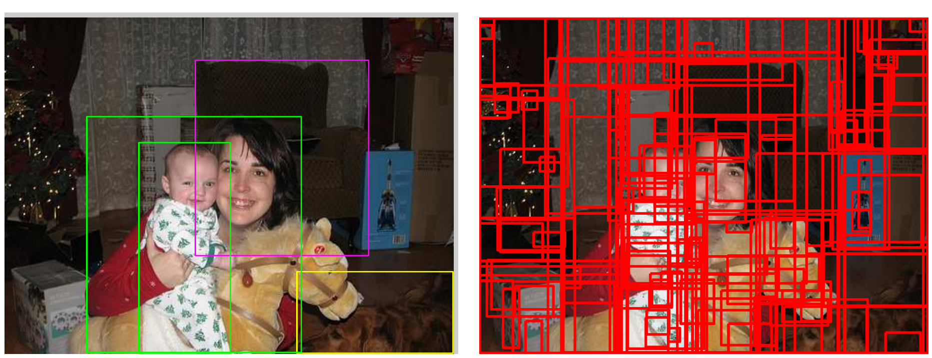

Figure 3. In the refinement stage, we formulate a multiple instance learning bag per image and bag instances correspond to each window proposals from selective search. Binary bag labels correspond to image-level annotations of whether the target object exists in the image or not. (Left) ground truth bounding boxes color coded with category labels. green: person, yellow: dog, and magenta: sofa, (Right) visualization of 100 random subset of window proposals.

We propose to use a smoothing technique studied by Nesterov (2005) for convex functions.

Nesterov's smoothing technique We only recall here the simpler form of Nesterov's results that is relevant for our purpose. Consider a non-smooth function that can be written in the following form:

$$g(\mathbf{w}) := \max_{\mathbf{u} \in \Delta} \langle \mathbf{A}\mathbf{w}, \mathbf{u} \rangle, \tag{10}$$

where $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{A}$ is in $\mathbb{R}^{m \times d}$, and Δ denotes the probability simplex, $\Delta = \{\mathbf{x} : \sum_{i=1}^m x_i = 1, x_i \geq 0\}$. Smoothing here consists of adding a strongly convex function ω in the maximization problem

$$g_\mu(\mathbf{w}) := \max_{\mathbf{u} \in \Delta} \Big[\langle \mathbf{A}\mathbf{w}, \mathbf{u} \rangle - \frac{\mu}{2}\omega(\mathbf{u}) \Big] . \tag{11}$$

The resulting function g_μ is differentiable for all $\mu > 0$, and its gradient is

$$\nabla g_\mu(\mathbf{w}) = \mathbf{A}^\intercal \mathbf{u}^\star(\mathbf{w}), \tag{12}$$

where $\mathbf{u}^\star(\mathbf{w})$ is the unique solution of (11). The parameter μ controls the amount of smoothing. Clearly, $g_\mu(\mathbf{w}) \to g(\mathbf{w})$ for all $\mathbf{w} \in \mathcal{W}$ as $\mu \to 0$. As Nesterov (2005) shows, for a given target approximation accuracy ϵ, there is an optimal amount of smoothing $\mu(\epsilon)$ that can be derived from a convex optimization perspective using the strong convexity parameter of $\omega(\cdot)$ on Δ and the (usually unknown) Lipschitz constant of g. In the experiments, we shall simply learn the parameter μ from data.

Smoothing the latent SVM We now apply Nesterov's smoothing technique to the latent SVM objective function. As we shall see, the smoothed objective takes a simple form, which can be efficiently computed in the latent SVM framework. Furthermore, smoothing latent SVM implicitly models uncertainty in the selection of the best configuration $\mathbf{z}$ in $\mathcal{Z}$, as shown by Kumar et al. (2012) for a different smoothing scheme.

In order to smooth the functions $f_{\mathbf{x}_i}$ defined in (9), we first notice that

$$f_{\mathbf{x}_i}(\mathbf{w}) = \max_{\mathbf{u} \in \Delta} \langle \mathbf{A}_{\mathbf{x}_i}\mathbf{w}, \mathbf{u} \rangle, \tag{13}$$

where $\mathbf{A}_{\mathbf{x}_i}$ is a matrix of size $|\mathcal{Z}| \times d$ such that the j-th row of $\mathbf{A}_{\mathbf{x}_i}$ is the feature vector $\phi(\mathbf{x}_i, \mathbf{z}_j)$ and $\mathbf{z}_j$ is the j-th element of $\mathcal{Z}$. Considering any strongly convex function ω and parameter $\mu > 0$, the smoothed latent SVM objective is obtained by replacing in (8)

- the functions $f_{\mathbf{x}_i}$ by their smoothed counterparts $f_{\mathbf{x}_i,\mu}$ obtained by applying (11) to (13);
- the non-smooth hinge-loss function l by any smooth loss.

Objective and gradient evaluations An important issue remains the computational tractability of the new formulation in terms of objective and gradient evaluations, in order to use quasi-Newton optimization techniques. The choice of the strongly convex function ω is crucial in this respect.

There are two functions known to be strongly convex on the simplex: i) the Euclidean norm, ii) the entropy. In the case of the Euclidean-norm $\omega(\mathbf{u}) = \|\mathbf{u}\|_2^2$, it turns out that the smoothed counterpart can be efficiently computed using a projection on the simplex, as shown below.

$$\mathbf{u}^\star(\mathbf{w}) = \arg\min_{\mathbf{u} \in \Delta} \left\| \frac{1}{\mu}\mathbf{A}\mathbf{w} - \mathbf{u} \right\|_2^2, \tag{14}$$

where $\mathbf{u}^\star(\mathbf{w})$ is the solution of (11). Computing $\mathbf{A}\mathbf{w}$ requires a priori $O(|\mathcal{Z}|d)$ operations. The projection can be computed in $O(|\mathcal{Z}|)$ (see, e.g., Bach et al., 2012). Once $\mathbf{u}^\star$ is obtained, computing the gradient requires $O(d\|\mathbf{u}^\star\|_0)$ operations, where $\|\mathbf{u}^\star\|_0$ is the number of non-zero entries in $\mathbf{u}^\star$.

When the set $\mathcal{Z}$ is large, these complexities can be improved by leveraging two properties. First, the projection on the simplex is known to produce sparse solutions, the smoothing parameter μ controlling the sparsity of $\mathbf{u}^\star$; second, the projection preserves the order of the variables. As a result, the following heuristic can be justified. Assume that for some $N < |\mathcal{Z}|$, we can obtain the top-N entries of $\mathbf{A}\mathbf{w}$ without exhaustively exploring $\mathcal{Z}$. Then, performing the projection on these reduced set of N variables yields a vector $\mathbf{u}'$ which can be shown to be optimal for the original problem (14) whenever $\|\mathbf{u}'\|_0 < N$. In other words, whenever N is large enough and μ small enough, computing the gradient of $f_{\mathbf{x}_i,\mu}$ can be done in $O(Nd)$ operations. We use this heuristic in all our experiments.

6. Experiments

We performed two sets of experiments, one on a multiple instance learning dataset (Andrews et al., 2003) and the other on the PASCAL VOC 2007 data (Everingham et al.). The first experiment was designed to compare the multiple instance learning bag classification performance of LSVM with Smooth LSVM (SLSVM). The second experiment evaluates detection accuracy (measured in average precision) of our framework in comparison to baselines.

6.1. Multiple instance learning datasets

We evaluated our method in Section 5 on standard multiple instance learning datasets (Andrews et al., 2003). For preprocessing, we centered each feature dimension and ℓ_2 normalize the data. For fair comparison with (Andrews et al., 2003), we use the same initialization, where the initial weight vector is obtained by training an SVM with all

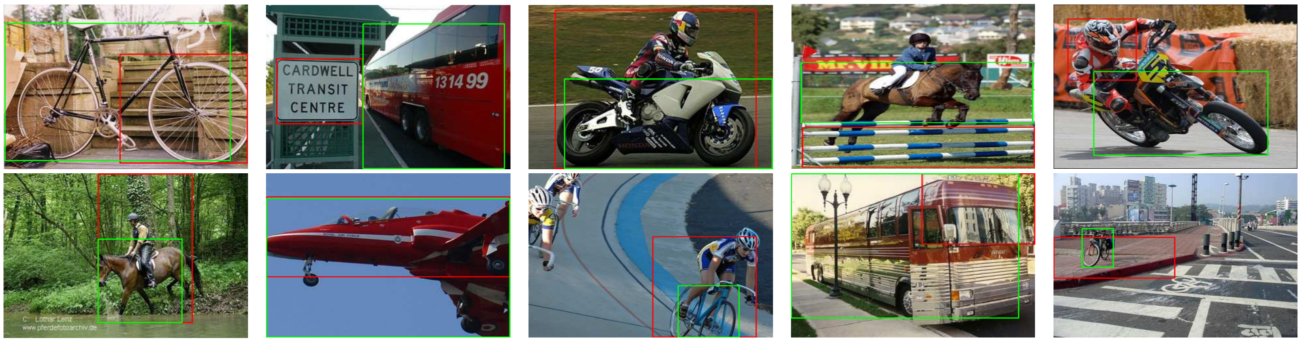

Figure 4. Visualization of some common failure cases of constructed positive windows by(Siva et al., 2012) vs our method. Red bounding boxes are constructed positive windows from (Siva et al., 2012). Green bounding boxes are constructed positive windows from our method.

Dataset	LSVM w/o bias	SLSVM w/o bias	LSVM w/ bias	SLSVM w/ bias
musk1	70.8 ± 14.4	80.3 ± 10.3	81.7 ± 14.5	79.2 ± 13.4
musk2	51.0 ± 10.9	79.5 ± 10.4	80.5 ± 9.9	84.3 ± 11.4
fox	51.5 ± 7.5	63.0 ± 11.8	57.0 ± 8.9	61.0 ± 12.6
elephant	81.5 ± 6.3	88.0 ± 6.7	81.5 ± 4.1	87.0 ± 6.3
tiger	79.5 ± 8.6	85.5 ± 6.4	86.0 ± 9.1	87.5 ± 7.9
trec1	94.3 ± 2.9	95.5 ± 2.6	95.3 ± 3.0	95.3 ± 2.8
trec2	69.0 ± 6.8	83.0 ± 6.5	86.5 ± 5.7	83.8 ± 7.4
trec3	77.5 ± 5.8	90.0 ± 5.8	85.5 ± 6.3	86.0 ± 6.5
trec4	77.3 ± 8.0	85.0 ± 5.1	85.3 ± 3.6	86.3 ± 5.2
trec7	74.5 ± 9.8	83.8 ± 4.0	82.5 ± 7.0	81.5 ± 5.8
trec9	66.8 ± 5.0	70.3 ± 5.7	68.8 ± 8.0	71.5 ± 6.4
trec10	71.0 ± 9.9	84.3 ± 5.4	80.8 ± 6.6	82.8 ± 7.3

Table 1. 10 fold average and standard deviation of the test accuracy on MIL dataset. The two methods start from the same initialization introduced in (Andrews et al., 2003)

Method	aeroplane		bicycle		boat		bus		horse		motorbike		mAP
	left	right	left	right	left	right	left	right	left	right	left	right	
(Deselaers et al., 2010)	9.1	23.6	33.4	49.4	0.0	0.0	0.0	16.4	9.6	9.1	20.9	16.1	16.0
(Pandey & Lazebnik, 2011)	7.5	21.1	38.5	44.8	0.3	0.5	0.0	0.3	45.9	17.3	43.8	27.2	20.8
(Deselaers et al., 2012)	5.3	18.1	48.6	61.6	0.0	0.0	0.0	16.4	29.1	14.1	47.7	16.2	21.4
(Russakovsky et al., 2012)	30.8		25.0		3.6		26.0		21.3		29.9		22.8
(Siva et al., 2012) with our features	23.2		15.4		5.1		2.0		6.2		17.4		11.6
Cover + SVM	23.4		43.5		8.1		33.9		24.7		40.2		29.0
Cover + LSVM	28.2		47.2		9.6		34.7		25.2		39.8		30.8

Table 2. Detection average precision (%) on PASCAL VOC 2007-6x2 test set. First three baseline methods report results limited to left and right subcategories of the objects.

VOC2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	pson	plant	sheep	sofa	train	tv	mAP
(Siva & Xiang, 2011)	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0	13.9
Cover + SVM	23.4	43.5	22.4	8.1	6.2	33.9	33.8	30.4	0.1	17.9	11.5	17.1	24.7	40.2	2.4	14.8	21.4	15.1	31.9	6.2	20.3
Cover + LSVM	28.2	47.2	17.6	9.6	6.5	34.7	35.5	31.5	0.3	21.7	13.2	20.7	25.2	39.8	12.6	18.6	21.2	18.6	31.7	10.2	22.2
Cover + SLSVM	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7

Table 3. Detection average precision (%) on full PASCAL VOC 2007 test set.

the negative instances and bag-averaged positive instances. For this experiment, we performed 10 fold cross validation on C and μ. Table 1 shows the experimental results. Without the bias, our method significantly performs better than LSVM method and with the bias, our method shows modest improvement in most cases.

6.2. Weakly-supervised object detection

To implement our weakly-supervised detection system we need suitable image features for computing the nearest neighbors of each image window in Section 4 and for learning object detectors. We use the recently proposed R-CNN (Girshick et al., 2014) detection framework to compute features on image windows in both cases. Specifically, we use the convolutional neural network (CNN) distributed with DeCAF (Donahue et al., 2014), which is trained on the ImageNet ILSVRC 2012 dataset (using only image-level annotations). We avoid using the better performing CNN that is fine-tuned on PASCAL data, as described in (Girshick et al., 2014), because fine-tuning requires instance-level annotations.

We report detection accuracy as average precision on the standard benchmark dataset for object detection, PASCAL VOC 2007 *test* (Everingham et al.). We compare to five different baseline methods that learn object detectors with limited annotations. Note that other baseline methods use additional information besides the one-bit image-level annotations. Deselaers et al. (2010; 2012) use a set of 799 images with bounding box annotations as meta-training data. In addition to bounding box annotations, Deselaers et al. (2010; 2012); Pandey & Lazebnik (2011) use extra instance level annotations such as *pose*, *difficult* and *truncated*. Siva et al. (2012); Russakovsky et al. (2012) use *difficult* instance annotations but not *pose* or *truncated*. First, we report the detection average precision on 6 subsets of classes in table 2 to compare with Deselaers et al. (2010; 2012); Pandey & Lazebnik (2011).

To evaluate the efficacy of our initialization, we compare it to the state-of-the-art algorithm recently proposed by (Siva et al., 2012). Their method constructs a set of positive windows by looping over each positive image and picking the instance that has the maximum distance to its nearest neighbor over all negative instances (and thus the name negative data *mining* algorithm). For a fair comparison, we used the same window proposals, the same features (Girshick et al., 2014), the same L2 distance metric, and the same PASCAL 2007 detection evaluation criteria. The class mean average precision for the mining algorithm was 11.6% compared to 29.0% obtained by our initialization procedure. Figure 4 visualizes some command failure modes in our implementation of (Siva et al., 2012). Since the negative mining method does not take into account the similarity among positive windows (in contrast to our method) our intuition is that the method is less robust to intra-class variations and background clutter. Therefore, it often latches onto background objects (i.e. hurdle in horse images, street signs in bus images), onto parts of the full objects (i.e. wheels of bicycles), or merges two different objects (i.e. rider and motorcycle). It is worth noting that Pandey & Lazebnik (2011); Siva et al. (2012) use the CorLoc metric[1] as the evaluation metric to report results on PASCAL *test* set. In contrast, in our experiments, we exactly follow the PASCAL VOC evaluation protocol (and use the PASCAL VOC devkit scoring software) and report detection average precision.

Table 3 shows the detection result on the full PASCAL 2007 dataset. There are two baseline methods (Siva & Xiang, 2011; Russakovsky et al., 2012) which report the result on the full dataset. Unfortunately, we were not able to obtain the per-class average precision data from the authors of (Russakovsky et al., 2012) except the class mean average precision (mAP) of 15.0%. As shown in Table 3, the initial detector model trained from the constructed set of positive windows already produces good object detectors but we can provide further improvement by optimizing the MIL objective.

7. Conclusion

We developed a framework for learning to localize objects with one-bit object presence labels. Our results show that the proposed framework can construct a set of positive windows to train initial detection models and improve the models with the refinement optimization method. We achieve state-of-the-art performance for object detection with minimal supervision on the standard benchmark object detection dataset. Source code will be available on the author's website.

Acknowledgement

We thank Yong Jae Lee for helpful insights and discussions. H. Song was supported by Samsung Scholarship Foundation. J. Mairal and Z. Harchaoui was funded by the INRIA-UC Berkeley associated team "Hyperion", a grant from the France-Berkeley fund, the Gargantua project under program Mastodons of CNRS, and the LabEx PERSYVAL-Lab (ANR-11-LABX-0025). This work was partially supported by ONR N00014-11-1-0688, NSF, DARPA, and Toyota.

References

Alexe, B., Deselaers, T., and Ferrari, V. Classcut for unsupervised class segmentation. In *ECCV*, 2010.

Andrews, S, Tsochantaridis, I, and Hofmann, T. Support vector

[1]CorLoc was proposed by (Deselaers et al., 2010) to evaluate the detection results on PASCAL *train* set

machines for multiple-instance learning. In *NIPS*, 2003.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

Barinova, O., Lempitsky, V., and Kohli, P. On detection of multiple object instances using hough transforms. *IEEE TPAMI*, 2012.

Boyd, S. P. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Chen, X., Shrivastava, A., and and, A. Gupta. Neil: Extracting visual knowledge from web data. In *ICCV*, 2013.

Chen, Y., Shioi, H., Montesinos, C. Fuentes, Koh, L. P., Wich, S., and Krause, A. Active detection via adaptive submodularity. In *ICML*, 2014.

Chum, O. and Zisserman, A. An exemplar model for learning object classes. In *CVPR*, 2007.

Crandall, D. and Huttenlocher, D. Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV*. 2006.

Darrell, T., Sclaroff, S., and Pentland, A. Segmentation by minimal description. In *ICCV*, 1990.

Deselaers, T., Alex, B., and Ferrari, V. Localizing objects while learning their appearance. In *ECCV*, 2010.

Deselaers, T., Alex, B., and Ferrari, V. Weakly supervised localization and learning with generic knowledge. *IJCV*, 2012.

Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. What makes paris look like paris? In *SIGGRAPH*, 2012.

Doersch, C., Gupta, A., and Efros, A. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *ICML*, 2014.

Endres, I., Shih, K., and Hoeim, D. Learning collections of part models for object recognition. In *CVPR*, 2013.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The PASCAL Visual Object Classes (VOC) Challenge. *IJCV*, 2010.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. Object detection with discriminatively trained part based models. *IEEE TPAMI*, 32(9), 2010.

Fergus, R., Perona, P., and Zisserman, A. Weakly supervised scale-invariant learning of models for visual recognition. *IJCV*, 2007.

Fukunaga, K. and Hostetler, L. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory*, 1975.

Galleguillos, C., Babenko, B., Rabinovich, A., and Belongie, S. Weakly supervised object localization with stable segmentations. In *ECCV*, 2008.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

Joulin, A. and Bach, F. A convex relaxation for weakly supervised classifiers. In *ICML*, 2012.

Joulin, A., Bach, F., and Ponce, J. Discriminative clustering for image co-segmentation. In *CVPR*, 2010.

Juneja, M., Vedaldi, A., Jawahar, V., and Zisserman, A. Blocks that shout: Distinctive parts for scene classification. In *CVPR*, 2013.

Kim, G., Xing, E.P., Fei-Fei, L., and Kanade, T. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011.

Kumar, P, Packer, B, and Koller, D. Modeling latent variable uncertainty for loss-based learning. In *ICML*, 2012.

Leibe, B., Leonardis, A., and Schiele, B. Combined object categorization and segmentation with an implicit chape model. In *ECCVW*, 2004.

Li, Y., Tsang, I., Kwok, J., and Zhou, Z. Convex and scalable weakly labeled svms. In *ICML*, 2013.

Long, P.M. and Tan, L. PAC learning axis aligned rectangles with respect to product distributions from multiple-instance examples. In *Proc. Comp. Learning Theory*, 1996.

Micolajczyk, K., Leibe, G., and Schiele, B. Multiple object class detection with a generative model. In *CVPR*, 2006.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1), 2005.

Nocedal, J. and Wright, S. *Numerical Optimization*. Springer, 1999.

Pandey, M. and Lazebnik, S. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

Raptis, M., Kokkinos, I., and Soatto, S. Discovering discriminative action parts from mid-level video representations. In *CVPR*, 2012.

Rother, C., Minka, T., Blake, A., and Kolmogorov, V. Cosegmentation of image pairs by histogram matching incorporating a global constraint into MRFs. In *CVPR*, 2006.

Russakovsky, O., Lin, Y., Yu, K., and Fei Fei, L. Object-centric spatial pooling for image classification. In *ECCV*, 2012.

Singh, S., Gupta, A., and Efros, A. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

Siva, P. and Xiang, T. Weakly supervised object detector learning with model drift detection. In *ICCV*, 2011.

Siva, P., Russell, C., and Xiang, T. In defence of negative mining for annotating weakly labelled data. In *ECCV*, 2012.

Uijlings, J., van de Sande, K., Gevers, T., and Smeulders, A. Selective search for object recognition. In *IJCV*, 2013.

Weber, M., Welling, M., and Perona, P. Towards automatic discovery of object categories. In *CVPR*, 2000a.

Weber, M., Welling, M., and Perona, P. Unsupervised learning of models for recognition. In *ECCV*, 2000b.

Wolsey, L. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2:385–393, 1982.

Yu, C.N. and Joachims, T. Learning structural svms with latent variables. In *ICML*, 2009.

Yuille, A.L. and Rangarajan, A. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.